Systems Software for the Next Generation of Storage

Horizontal Scaling Solution using Linux Environment

December 14, 2001

Carter George
Vice President, Corporate Development
PolyServe, Inc.

PolyServe

- **Goal**: Build a new layer of systems software for thin, dense, and rackmount servers
- **Provide**: the systems software to allow data centers to "scale out"—enable horizontal scaling
- Build the best possible SAN file system as the core technology
  - Significant differentiation from other "cluster" file systems
  - Portable across many OS's
  - Scalable to large node counts
  - Integrates with Oracle9i RAC and DB2 EEE

Example: Dell Versus Sun UE

- 36 Dell 2 proc servers
- 72 processors total
- Processors are 1.2 GHz
- 144 GB Memory
- 36 Qlogic 2200 FC HBA
- 72 Gigabit Ethernets
- Linux 7.1

- **Total = ~$300,000**

- 1 Sun UE 15000
- 72 processors
- Processors are 900 MHz
- 288 GB Memory
- Gigabit Ethernet
- Solaris

- **Total = ~$4,000,000**

The Intel-based configuration has more performance and more I/O bandwidth at 1/10th the cost….

PolyServe

Benefits of Horizontal Scaling

- *Extremely* Compelling Economics
- Granularity of **Scaling**
  - Grow servers and storage independently
  - If load grows, add more servers—no forklift upgrades!
- Granularity of **Availability**
  - Server failure ⇒ only 1/nth of processing capacity is lost
  - All servers are active
  - With 2 large servers, a server failure costs 50%
  - Better availability, rolling upgrades, planned maintenance all possible
- Flexibility
  - Can deploy standby servers into application pools based on QoS metrics, SLAs, etc…

The PolyServe Product Line

- PolyServe Application Manager
  - Next generation clustering for thin and dense servers
  - Designed for rack-optimized, high-node-count architectures
  - Large node counts, no dependency on shared disk
  - Start, stop, monitor, and move applications in server farms
  - Runs identically on Windows, Linux, and Solaris
  - Shipping already – many customers in Japan!
- PolyServe Matrix Server
  - Includes all of PolyServe Application Manager
  - PolyServe SAN File System
  - Breakthrough shared file system software for storage
  - Full support for horizontal scaling with Oracle9i RAC
  - Linux beta this quarter, GA 1Q02; Windows & Solaris to follow

PolyServe

Analysts…

•Distributed clustered file systems are rapidly becoming real. Connectivity mediums like InfiniBand are the great enablers for these types of technologies.

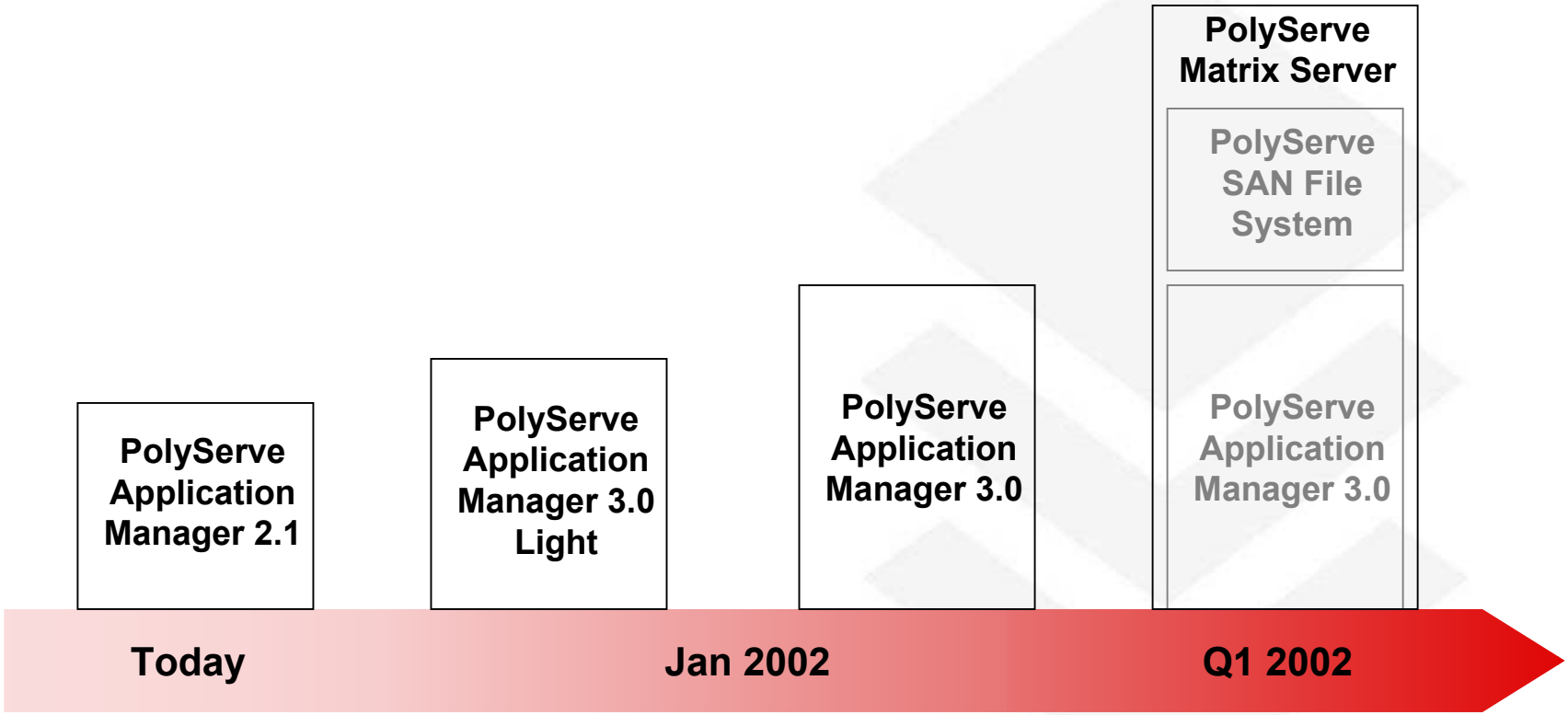
•We view the combination as a killer to enable low cost servers to very tightly couple over a super low latency interconnect and form massively scaleable systems. This will be the way of the future.

•PolyServe is delivering on the dream of true clustered file systems. The Oracle*9i* integration is a perfect application for this technology, and users will see an immediate benefit."

•Steve Duplessie, Enterprise Storage Group

PolyServe

PolyServe Roadmap



PolyServe

PolyServe Application Manager

- High Availability
  - Start, stop, move, and failover applications
- Low Cost of Ownership
  - Low acquisition price
  - Extremely easy to install and configure – 15 minute install time
  - Extremely easy to manage – NOT like Unix clusters
- Support for many nodes
  - Up to 64 nodes
  - Tested up to 100 nodes
  - N:1 and N:m failover scenarios
- Understand the Internet Data Center Usage Model
  - Not just database applications
  - Does not assume shared storage
- Data replication engine
- Extensible scripts and monitors

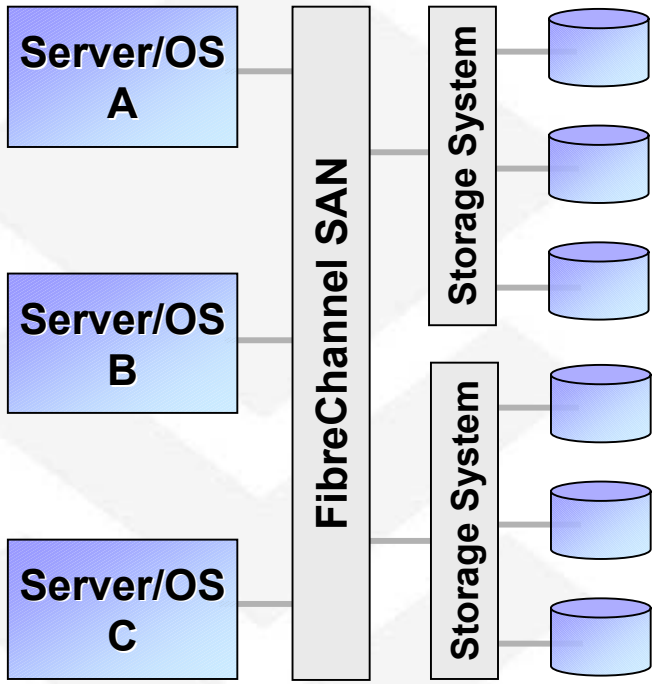
PolyServe

PolyServe Matrix Server: Design Goals

Symmetric Cluster File System

Complete concurrent data sharing across many servers with:

1. Data Integrity
2. High Availability
3. Recoverability
4. Scalability
5. Performance
6. Ease of Use
7. Portable: Linux, Windows, Unix

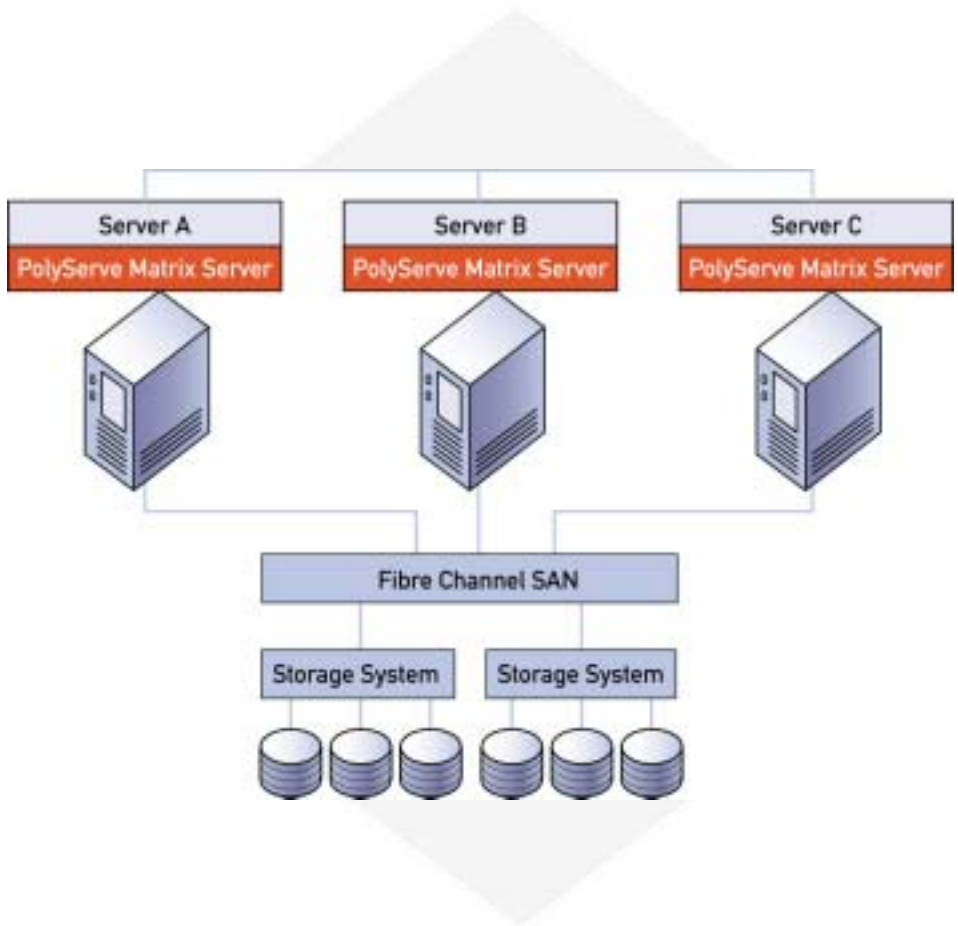


Matrix Server redefines the way servers talk to storage

PolyServe

PolyServe Matrix Server Architecture

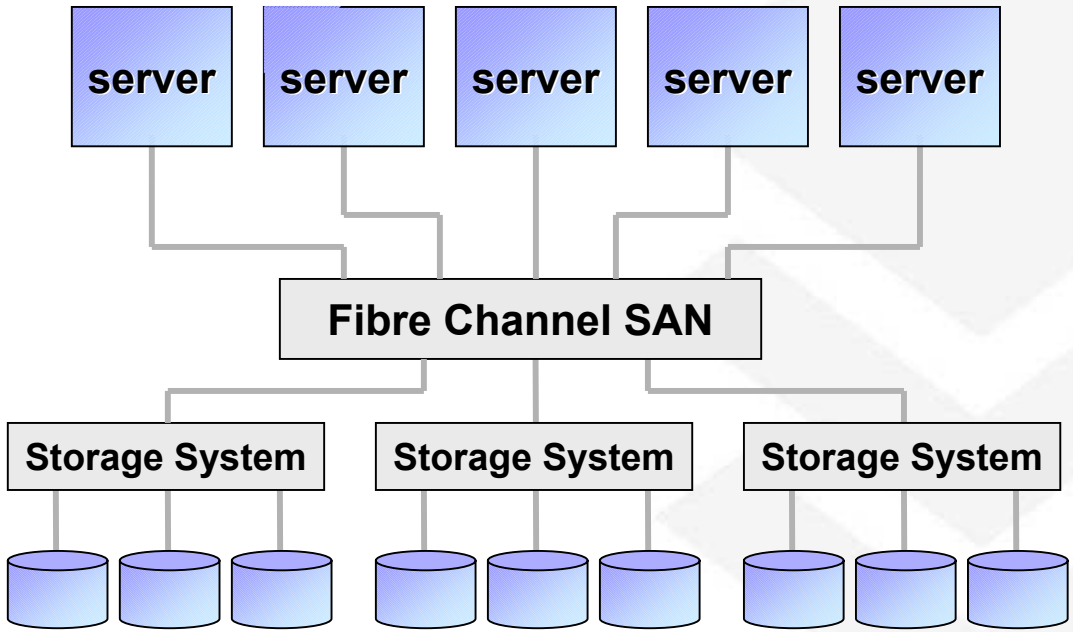
- No single point of failure
- Completely Symmetric
- Allows dynamic addition of servers and disk subsystems
- Provides coordinated access to shared resources
  - **SAN File System**
  - **Distributed Lock Manager**
  - **Distributed Collaborative Cache (Cache Fusion)**
- Unsurpassed data integrity and availability
- On-line recovery without interruption
- Single point of administration
- Works with FC, Gig-E, IB



PolyServe

2. MxS Availability: Adding Resources Online

Servers can be added dynamically to handle increasing application load



Storage can be added dynamically (and independently) as storage demands grow

PolyServe

3. Matrix Server Recoverability: File System Recovery

- Online Recovery of File System
- Data remains accessible to all live servers if a server or network fails
- Cascading recovery
  - Recovery fails over if multiple servers fail
  - Multiple simultaneous failures
  - Cascading failures
- MxS is designed to provide 24 x 7 access to the SAN file system under even catastrophic failure conditions

PolyServe

5. Matrix Server: File System Performance

- High-speed Distributed Lock Manager
- Distributed Collaborative Cache (Cache Fusion)
  - Supra linear read performance
  - Requires low latency interconnect (eg. IB)
  - Otherwise, copyback cache coherence
  - Note: Copyback better than write-through
- Optimized on-disk file and data structure layout
  - No legacy format to protect
  - Native design center: high degree of concurrency
  - No false sharing

4. Matrix Server: Scalability

- Designed for large node counts
  - 16 to 32 with IP interconnects
  - Number dependent on contention and workload
  - Many more with Infiniband (256?)
- Support for multiple node to node interconnects
  - Gigabit Ethernet
  - Infiniband
  - VI over Fibre Channel

Matrix Server and 9i RAC

- **Matrix Server is *the* file system to support 9i RAC implementation in large Data Centers**
  - **Best recoverability**
  - **Best scalability**
  - **Best TCO and cost of management**
- **Matrix Server will have full Oracle ODM support**
- **Matrix Server with ODM and 9i RAC may outperform raw disk**
- **Matrix Server with 9i RAC will outperform large Unix systems at 64 processors**
- **Matrix Server is designed with horizontal scaling and Oracle in mind**

PolyServe

Matrix Server's Value-Add for Oracle

- **Manageability** of data files
  - Customers don't want raw devices—as they are used to
  - See META Group analysis
- **Manageability** of **Oracle Home**
  - 105,000 files in each fully-installed home
  - Intractable for many-node installations without infrastructure
- Integration with other processes
  - ETL scripts can run across multiple nodes
  - Bulk load etc
- **Better performance**
  - ODM implementation: demonstrated 39%, 19% CPU utilization reduction for DBWR, LGWR
  - Context switch reduction of 6% (important on Intel)

What the analysts are saying…

•However, the most exciting thing about 9i - and by far the greatest technical challenge - is Real Application Clusters (RAC), a major extension of Oracle Parallel Server that promises the ability to scale out a database by adding new logical nodes in a shared environment on any of the platforms that Oracle supports…

•However, we do not expect RAC to work immediately without a hiccup in production systems. We expect that Oracle will need to bring out intermediate upgrades, and that it may not work outside very specialized applications on any of the platforms until clustered file systems are available for those major platforms….

•META Group

PolyServe

Four Keys to Horizontal Scaling

- *Why Haven't We Been Doing this for Years?*

- Four technology advances have only now made this possible:

- New Server Interconnects
- Storage Networks
- Distributed Databases
- SAN File System

- *PolyServe provides the last missing piece*

PolyServe

PolyServe

- Well-Funded Startup
  - ~$25 Million raised post-correction – November 2000
  - Top-tier VC's: Greylock, NEA
  - 2 Years in development with revenue today (rack-optimized high-availability product)
- Top Engineering Team
  - Core of former Sequent NUMA OS kernel group; senior engineers from Veritas, Microsoft, Sun, IBM, etc.
  - Ten years' track record delivering mission-critical enterprise software
- Strong Partnerships
  - Oracle, IBM, Brocade, EMC
- Target Market = Major Customers
  - Fortune 500 Telcos, Banks, Manufacturers, Technology Companies

What Some Analysts Are Saying….

- Getting to effective data sharing requires the creation of high-performance sharable file systems for networked storage. These file systems define how data is stored in a networked storage environment, as well as contain the rules by which any host can access, retrieve, and manipulate the data.

- As such, file systems for networked storage are critical enablers of heterogeneous data sharing. We believe that PolyServe shares this vision and will deliver unique data sharing products.

- John Webster, Illuminata

Questions to Ask About Other Clustered File Systems

- Symmetric Architecture
  - No Central Locking or Metadata
- Distributed Lock Manager
- Cache
  - Cache Coherent
  - Cache Fusion
- Online Recovery
  - Multiple failures
  - Cascading failures
- Database Support
  - Oracle9i RAC with ODM
  - Sybase, DB2, SQL Server
- Operating System Support
  - Linux
  - Windows
  - Unix
- Performance
  - On disk layout
  - Number of nodes
  - Multiple concurrent writers
- Availability
  - Add disks and servers online
  - Built-in failover
  - Full app and dev monitoring
  - Online resize

Competition: Some Examples

- **Veritas Cluster File System**
  - Legacy on-disk format
  - No ODM or Quick I/O
  - Central Metadata bottleneck
  - Solaris only – no Windows port
  - Linux port? Who knows
  - Poor scaling: 3 or 4 nodes max
  - No Cache Fusion

- **Sistina GFS**
  - Central locking bottleneck
  - No DLM
  - No online recovery
  - Linux only - No Windows port
  - Write-through cache
  - No Cache fusion

Summary

- PolyServe: Serious commercial software for serious commercial applications
  - Best Linux High-Availability Software (PolyServe Application Manager)
  - Best Linux SAN File System (PolyServe Matrix Server)
  - Best Oracle story (Full Oracle ODM and 9i RAC support)
  - Best roadmap (Full Windows and Solaris support in 2002)
  - Best systems software for horizontal scaling
  - HUGE economic benefits for major IT data centers!
- NTT Comware is exclusive PolyServe Partner in Japan

PolyServe

Systems Software for the Next Generation of Storage

Thank you